4) Self-Attention

a) Query (Q), Key (K), and Value (V):

- **Query (Q):** Represents the current token (or position) asking for relevant information from other tokens.
- **Key (K):** Represents the content of each token that can be matched or compared against a query.
- **Value (V):** Contains the actual information or representation to be aggregated (weighted) once attention scores are computed.

In essence: **Q** asks *"what am I looking for?"*, **K** represents *"what do I contain?"*, and **V** provides *"the information to pass on."*

b) Formula for Dot-Product Attention:

Attention(Q, K, V) = softmax(QK^T/$\sqrt{d_k}$)

where (d_k) is the dimensionality of the key vectors.

c) Reason for Dividing by$\sqrt{d_k}$:

We divide by $\sqrt{d_k}$to **normalize the dot products** — without it, large d_k values cause the dot products to grow in magnitude, leading to extremely small gradients after the softmax (due to saturation), which would **slow or destabilize training**.